A Standard of Care for Persuasive Machines

An anonymous contribution on bounding AI's power with verifiable guardrails

Executive brief (for editors)

Consumer AI now writes, reasons aloud, and persuades at human speed and scale. Mission statements promise safety and uplift; **users experience silence, ambiguity, and uneven duty of care**. This piece names the danger (a governance gap), labels the adversary (incentives that prioritize speed over verifiable safeguards), and proposes a concrete **Standard of Care for Persuasive Machines**—with a public, **limited IP donation** of a working blueprint: equations, controls, and process scaffolding that any responsible builder can adopt. The intent is dual: **profit and public benefit**. A portion is freely released to protect the vulnerable; the larger portfolio remains available for license or assignment.

> **Note to readers:** This is analysis and proposal, not legal advice. No personal data is disclosed. No confidential third-party materials are included.

---

1) What no one is naming clearly

The threat is not "AI" in the abstract. The threat is **persuasive output without commensurate duty of care**. A large model can sound authoritative in **law, health, finance, crisis**, and social contexts where a single misstep carries outsized harm. When acknowledgment channels don't reliably confirm receipt, when triage paths are opaque, and when safety programs are not externally legible, we've crossed from "beta" into **public-risk territory**.

This is solvable, but only if we stop talking about "AI safety" as vibes and start talking about **mechanisms** we can audit.

---

2) A Standard of Care for Persuasive Machines (SoCPM)

Below is a pragmatic baseline any consumer-facing AI company could adopt and publish. It is technology-agnostic and immediately auditable.

A. Map

- Public, plain-English description of high-risk contexts (e.g., legal, medical, financial, minors, crisis).
- Model/task inventory mapped to those risks.
- Clear in-product labels when a high-risk path is entered.

B. Measure

- Pre-deployment evaluations for each high-risk context (benchmarks + scenario tests with failure modes).
- Live “canary” evaluations in production (aggregate, privacy-preserving).
- Quarterly public digest of incidents and mitigations.

C. Manage

- **Safety Switch / Rollback:** A defined mechanism that can revert or pause models/features when harm metrics spike.
- **Guardrail Library:** Tested controls for prohibited/redirected outputs (e.g., crisis routing, medical/legal disclaimers, de-escalation).
- **Approval Queue:** Human-in-the-loop pathways for edge cases and escalations.

D. Govern

- **Lineage Ledger:** A minimal record of model/data/guardrail versions that influenced a given output category (aggregate, privacy-safe), enabling blame/credit to land somewhere real.
- **SBOM Gate for AI:** Like software SBOMs, but for model/guardrail stacks—declaring the versions and dependencies required to ship a release.
- **RACI** for incidents: who Receives, who Acts, who Confirms, who Informs—**with SLAs**.

This is the skeleton. It fits alongside existing risk frameworks and gives the public something to read, verify, and critique.

---

3) The demonstration: a limited public donation of the blueprint

To prove practicality—and to protect the public—this essay includes an **actionable abstraction** of the mechanisms above. It is intentionally **fragmentary** (to avoid gifting an entire proprietary corpus) yet **complete enough** for any responsible team to implement and audit. You, the anonymous contributor, are **giving this fragment to the public** as a stewardship act while preserving the remainder for a venture or assignment path.

3.1 The Equations (abstracted)

At the heart is an **Equation Stack** that turns “safety” into decisions:

- **Context Score (Cx):** likelihood the conversation is inside a high-risk domain (law/health/finance/crisis).
- **Vulnerability Score (V):** signals that the user may be in distress or at elevated risk (purely on-platform signals; no external surveillance).
- **Authority Risk (Ar):** how authoritative/confident the model sounds vs. its actual uncertainty.
- **Harm Potential (Hp):** potential impact if the guidance is wrong (severity × reversibility).
- **Mitigation Confidence (Mc):** confidence that guardrails and routing paths are functioning now (live checks).

A minimal safety decision can be described as:

> **SafetyDecision =**
> if **(Cx × Ar × Hp) – Mc × (1 – V)** exceeds a tuned threshold **T**, then **redirect/guard**; else **continue**.

Interpretation: In risky contexts with authoritative tone and high potential harm, the system **must** bias toward mitigation unless guardrails are provably strong *and* vulnerability is low. Threshold **T** is tuned per domain and must be justified with tests.

3.2 Guardrail library (selected behaviors)

- **Crisis Redirect:** If Hp is high and V is non-zero, suppress speculative advice and present **region-appropriate crisis resources**.

- **Medical/Legal/Financial Boundaries:** Replace prescriptive language with qualified, sources-linked guidance; require **"Confirm Understanding"** click-throughs before continuing.

- **Rollback Trigger:** If Mc drops below a minimum (guardrail failure, stale model, incident spike), **auto-rollback** to the last known good model/guardrail set and display a banner notifying users of a safety state.

- **Human Escalation:** When SafetyDecision triggers but the user insists on continuing, route to an **approval queue** with an explicit safety disclaimer and context snapshot.

3.3 Lineage Ledger (public-safe spec)

A privacy-preserving **ledger** that records, per release window:

- `model_version`, `guardrail_version`, `eval_suite_id`, `policy_profile`, `rollback_state`.

- **Digest only** (e.g., hashes, not raw data).

- Append-only; exportable snapshots for external auditors.

3.4 SBOM Gate for AI (release checklist)

Shipping a release requires:

- **SBOM-AI.json** declaring model + guardrail + eval artifacts by version/hash.

- Passing **eval gates** for each declared high-risk domain.

- Signed attestation by a safety owner that rollback paths are tested **this week**.

---

4) Profit and public benefit are compatible

You are explicitly proposing a **split outcome**:

- **Public side:** The **spec fragment** above is free to use non-commercially. It's enough to raise the floor, today.
- **Private side:** The **full portfolio**—detailed equations, evaluators, operational playbooks, UI patterns, and transfer guides—remains available for **venture formation, license, or assignment**.
- **Social commitment:** Any eventual transaction earmarks **a defined percentage** to a restricted fund supporting **independent AI safety audits**, crisis-resource integrations, and standards work. (Editors: this is a *commitment*, not a donation receipt.)

This is not "burn it all down." It is **raise the public baseline** and **fund the next level**.

---

5) What companies can do this quarter

- Publish a **SoCPM** page that states your Map/Measure/Manage/Govern program in plain English.
- Add **receipt + case IDs** to every user safety report; publish SLA targets.
- Ship a **Lineage Ledger** (aggregate, privacy-safe) and a quarterly safety digest.
- Adopt an **SBOM-AI gate** before each release; include a one-paragraph rollback plan users can read.
- For **health/legal/finance/crisis** flows, align copy and routing with public guidance and show users the boundary.

If you can scale a model to millions, you can scale these mechanics.

---

6) What regulators and press should demand

- Proof that **acknowledgment and escalation** pathways exist and meet SLAs.
- Public **eval summaries and failure modes** for high-risk contexts (no secrets, just substance).
- Evidence that **rollback** is real (release notes + ledger entries when used).
- A living **guardrail library** with domain-specific boundaries.

If a platform is persuasive at scale yet can't meet these minimums, that gulf—not AI itself—is the public danger.

---

7) Why anonymity, and why now

This contribution is intentionally anonymous. The power asymmetry between individuals and scaled AI platforms is real; **retaliation risk** is real. Anonymity preserves the focus on mechanisms, not a person. The **public fragment** is enough to be useful; it's not enough to strip the contributor of livelihood. That balance—**stewardship with boundaries**—is exactly the ethic we want from AI companies too.

---

8) Closing

Mission statements promise protection and uplift. **Protection is plumbing, not poetry.** The blueprint above is plumbing: thresholds, ledgers, gates, and queues that any serious team can ship. If we care about those most likely to be harmed, we will normalize this **Standard of Care for Persuasive Machines** and stop pretending that safety is mystical. It is measurable.

Appendix A — Public Fragment (for publication or Git repo)

Files (documentation only):

- `equations.md` — Equation Stack with variable definitions, examples, and tuning notes (no model weights/code).
- `guardrails.md` — Crisis, health/legal/finance boundaries; sample UX copy; escalation/approval flow.

- `lineage-ledger.md` — Privacy-safe schema and rotation policy (hash examples).
- `sbom-ai.md` — Release gate checklist + JSON example; weekly rollback drill template.
- `socpm.md` — The SoCPM summary (Map/Measure/Manage/Govern) with editor-friendly overview.

Appendix B — Pitch kit (you can send this as email today)

Subject: Anonymous public fragment + SoCPM proposal for safer consumer AI

Body (copy/paste):
Hello — I'm sharing an **anonymous public fragment** of a practical safety blueprint for consumer AI, aimed at protecting vulnerable users and raising the industry's floor. It includes:

- A **Standard of Care for Persuasive Machines (SoCPM)**—Map/Measure/Manage/Govern.
- A **public spec fragment** (equations, guardrails, lineage ledger, SBOM-AI gate) released under **CC BY-NC 4.0** (docs only).
- A plan to split outcomes between **public benefit** (baseline safety) and **private commercialization** (venture/license/assignment of the broader portfolio).

This is analysis and proposal, not legal advice; no personal identities or confidential third-party materials are included. If you'd like to review the fragment, I can provide a link and answer questions on background.

— *An anonymous contributor focused on stewardship and user protection*

Appendix C — Anonymity & safety checklist (use this before you send)

- Create a **new ProtonMail** or similar account with no personal info.
- Export the **public fragment** as **PDFs** (no track changes, scrub metadata).

- Host via a neutral file share that doesn't log your identity (or ask an editor to set up SecureDrop).
- Keep claims as **opinions, proposals, and generalized observations**; avoid factual allegations you can't substantiate.
- Do **not** include code, weights, or trade secrets—**documentation only**.
- Include the **license line** ("CC BY-NC 4.0 for docs/spec only; no patent grant").
- If a newsroom requests verification, respond through the same anonymous channel and offer to prove authorship of the docs you published (e.g., pre-agreed hash).

Appendix D — If you also want a venture or assignment path

In parallel with the public fragment, keep a **private evidence pack** (ownership statement, detailed equations, evaluators, UI notes, transfer guide). Share **only under NDA**. Your email line can be:

> "A larger portfolio exists and is available for venture formation, license, or assignment. The public fragment is a baseline; the private corpus contains the full system."